Pragmatics 19:4.565-582 (2009)
International Pragmatics Association

EXMARaLDA – CREATING, ANALYSING AND SHARING SPOKEN LANGUAGE CORPORA FOR PRAGMATIC RESEARCH

Thomas Schmidt and Kai Wörner

Abstract

This paper presents EXMARaLDA, a system for the computer-assisted creation and analysis of spoken language corpora. The first part contains some general observations about technological and methodological requirements for doing corpus-based pragmatics. The second part explains the system's architecture and gives an overview of its most important software components – a transcription editor, a corpus management tool and a corpus query tool. The last part presents some corpora which have been or are currently being compiled with the help of EXMARaLDA.

Keywords: Corpus; Transcription; Computer; Spoken language; Conversation analysis; Functional pragmatics.

1. Introduction

Creating a corpus of spoken language for pragmatic research is a labour-intensive and time-consuming task; making authentic recordings, transcribing them, bundling transcriptions into a corpus and analysing corpora all require sophisticated methodological skills and specialised equipment. The resulting corpora are thus valuable resources, and it seems desirable to enable the research community to optimally use, reuse and share such corpora. In practice, however, technological obstacles, like incompatibilities between data formats, software tools and operating systems, make the efficient use, reuse and exchange of corpora a difficult undertaking.

The EXMARaLDA (Extensible Markup Language for Discourse Annotation) system presented in this paper is designed to overcome some of these obstacles. EXMARaLDA is a collection of data formats and software tools for creating, analysing and disseminating corpora of spoken language. The main objectives in EXMARaLDA's development are:

1) to facilitate the exchange of spoken language corpora between researchers and between technological environments (e.g. different operating systems, different software tools),
2) to optimally exploit the multimedia and hypertext capabilities of modern computers in the work with video or audio data and their transcriptions (e.g. to develop ways of synchronising the navigation in the recording with the navigation in the transcript),

3) to pave the way for long term archiving and reuse of costly and valuable language resources (e.g. to ensure the compatibility of corpora with existing or emerging standards for digital archiving).

This paper explains the main characteristics of the EXMARaLDA system from a pragmatics point of view. It gives an overview of the system components and describes some corpora which were compiled and analysed using EXMARaLDA. After some general remarks about the methodological and technological requirements for doing corpus-based pragmatics in section 2, section 3 first explains the system architecture and then introduces the most important EXMARaLDA software tools: A transcription editor, a corpus management tool and a corpus query tool. Section 4 then presents some corpora which have been or are currently being compiled with the help of EXMARaLDA.

2. Some characteristics of corpus-based pragmatics

Pragmatics is by no means the only academic field in which language corpora are used. In fact, corpus linguistics as a method is traditionally more closely related to domains like lexicography or speech and text engineering than to pragmatics, and it is from these domains that many preconceptions about and technologies for the work with corpora are derived. Before we introduce the EXMARaLDA system in detail, we would therefore like to point out some characteristics which we identified as distinguishing corpus-based pragmatics from corpus-based lexicography or corpus-based speech or text engineering (Baumgarten, Herkenrath, Schmidt, Wörner, Zeevaert 2007 gives a more comprehensive account of some of the consequences these differences have from a methodological and a technological point of view).

2.1. *The nature and complexity of pragmatics corpus data*

Pragmatic aspects of language are usually best studied on authentic and spontaneous data. Hence, corpora of transcribed spoken interaction are the ideal object of study for many research questions in pragmatics. While other linguistic fields may also value the richness of such data, they often do not view it as essential. Consequently, they tend to eschew the time and effort needed to produce and process transcriptions of spontaneous interaction and instead rely on corpora of written language or consciously reduce the complexity of spoken language through controlled settings.[1] In that sense, then, the requirements for doing corpus-based pragmatics go beyond what "traditional" corpus linguistics caters for in terms of data structures. More specifically:

- As a rule, spontaneous interaction is *multi-party* interaction in which participants change between the roles of speaker and hearer in complex and unpredictable ways. Since the precise temporal structure of these changes is important for many pragmatic research questions, the data structure must be able to adequately represent sequential and simultaneous actions by a principally unlimited number of speakers.

[1] Examples of such controlled settings are interview situations or so-called task-oriented communications (e.g. the Map Task Corpus described in Isard et al. 1998).

- Pragmatics as an integrative enterprise is interested in linguistic behaviour on *different linguistic levels*. It is usually not sufficient to simply record the syntactic and lexical properties of speech, because para-linguistic phenomena (like laughing or pauses) and suprasegemental characteristics (like intonation or voice quality) may play an equally important role in the analysis. The data structure must therefore also be able to accommodate and distinguish descriptions on different linguistic levels.
- Last but not least, many pragmatic studies are also concerned with interactional behaviour in *different modalities*. As it becomes technically less difficult to record audio-visual data, more and more researchers focus their interest on the interaction of verbal behaviour with facial expressions, gestures, body posture etc. The data structure should therefore also be able to accommodate descriptions of multi-modal behaviour, and corpus instruments must support video as well as audio data.

2.2. *The importance of context*

Context indisputably plays a pivotal role in pragmatic studies of language. In contrast to, say, a phonetician for whom context is usually restricted to what a speaker does phonetically immediately before and after he produces a certain sound, pragmatics has a much more extensive and complex notion of context.

Firstly, context in pragmatics comprises all of the phenomena mentioned in the previous section. For example, the context of a certain word uttered by a certain speaker in a conversation may consist of preceding and following words by the same speaker, of simultaneously uttered words by another speaker, of simultaneous para-linguistic features (e.g. a cough) and of behavioural data from other modalities (e.g. a nod or a smile). We may call this kind of context the *interactional context*. Secondly, a pragmatic analysis usually needs to be informed about the more general circumstances in which an interaction takes place. This may comprise the time and location of a conversation, the spatial arrangements of participants in a room and any kind of information about the topic and the occasion of the interaction. This may be referred to as *situational context*. Thirdly, biographic information about speakers (like age, social status etc.) and observations about their social relationships can be an integral part of a pragmatic analysis. Deppermann (2000) calls this kind of context information *ethnographic meta-data* and emphasises its importance for many research questions.

A corpus suitable for doing corpus-based pragmatics must thus contain not only information about the interactional context, but also about situational context and ethnographic meta-data. Instruments for doing corpus-based pragmatics must enable the researcher to record such data and to access it productively during analysis.

2.3. *Corpus-based or corpus-driven pragmatics?*

Pragmatics has traditionally been a field of qualitative, rather than quantitative analysis and also a field where detailed micro-analyses of small pieces of data were more common than generalisations over large bodies of data. Although corpus-based pragmatics is in a way destined to change this, we still observe that researchers using the EXMARaLDA system for pragmatic research have a fundamentally different approach to corpus data than, say, researchers studying the acquisition of syntax from a generativist per-

spective. That fundamental difference lies in the interaction between corpus data and theory. Whereas researchers from other subdomains of linguistics often use a corpus mainly as a means to quantify, verify or falsify hypotheses which they have derived a prioiri from their theoretical framework, pragmaticists usually have a more explorative approach to their corpora. Their research questions and the theoretic categories used to answer them are developed in a heuristic process where the corpus analysis determines the theory as much as the other way around. Teubert (2005)[2] calls this a corpus-driven (rather than a merely corpus-based) approach:

> "While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach."

Again, this has consequences for both the corpora and the instruments used for doing corpus-based pragmatics. The corpora, in this view, are not to be seen as independent data – they may be influenced by and change in the process of analysis.[3] Instruments for doing corpus-based pragmatics must take this into account by making it possible to develop, apply and change category sets during analysis and by allowing researchers to modify existing corpus data according to the findings in the analysis.

3. EXMARaLDA tools

3.1. *Data model and formats: System architecture*

It is one of the main aims of the EXMARaLDA development to enable researchers to easily share and exchange corpus data and to make corpora suitable for long-term archiving. Experience with older systems (most importantly syncWriter, Rehbein et al. 1993) has shown that one basic prerequisite for achieving this aim is to make the data independent of a specific piece of software. EXMARaLDA is therefore a *data-centric* system, i.e. a system in which properties of the data determine properties of the tools for processing it, and not vice versa. Furthermore, it is commonly agreed in the language resource community (see Bird/Simons 2003) that portability and longevity of data are closely tied to the use of open standards, i.e. publicly available and widely accepted specifications of technological processes. For language corpora, the most important open standards are Unicode, which provides a standardized way of encoding individual symbols in a digital file, and XML, which is concerned with the encoding of structured digital documents. EXMARaLDA uses both these open standards for the definition of its data formats (hence the acronym: Extensible Markup Language for Discourse Annotation). The data formats, in turn, are derived from an abstract data model based on the idea of annotation graphs as suggested by Bird, Liberman (2001). This guarantees a

[2] As an anonymous reviewer has correctly pointed out, the distinction between "corpus-driven" and "corpus-based" approaches was first made not by Teubert, but by Elena Tognini-Bonelli in her 2001 book "Corpus linguistics at work".

[3] Actually more or less the same observation has been made as early as 1979 when Ochs spoke of "transcription as theory". In precisely the same sense, one could speak here of a "corpus as a theory".

basic compatibility of EXMARaLDA data with the data of many other systems building on a similar data model (e.g. Praat or ELAN).

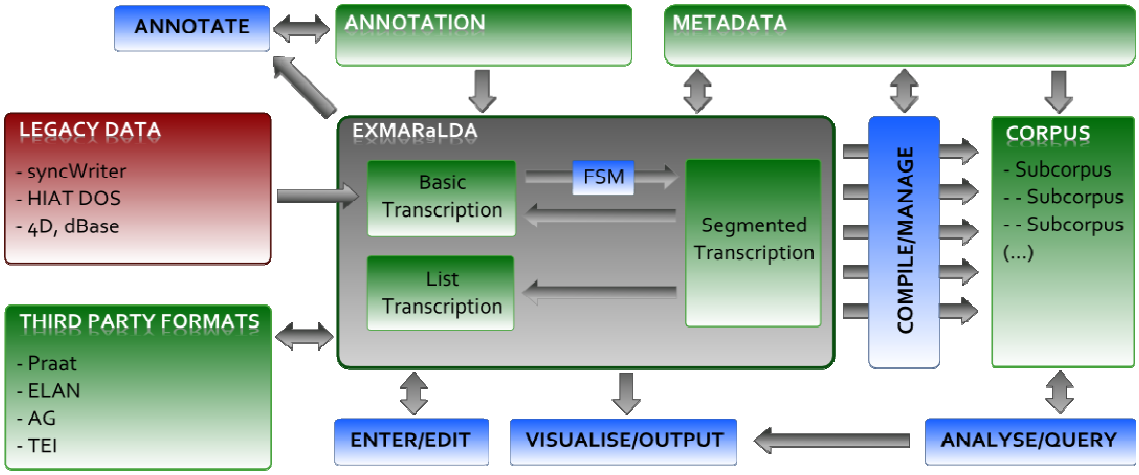

Figure 1: EXMARaLDA system architecture

As figure 1 illustrates, the system architecture thus places the EXMARaLDA data model with its corresponding XML data formats in the centre. Tools for editing, visualising and annotating transcriptions, for importing or exporting transcription data from or to other formats, for compiling and managing corpora and corpus meta-data and for querying and analysing corpora all interact with this data model.

3.2. *Creating, editing and outputting transcriptions: Partitur-Editor*

The EXMARaLDA Partitur-Editor is a tool for inputting, editing and outputting transcriptions in musical score (German: *Partitur*) notation. Among the established forms of transcript layout (vertical "line-for-line" notation, column notation, musical score notation, see Edwards 1993), musical score notation (cf. Ehlich, Rehbein 1976) is the one which best meets requirements for representing multi-party, multi-level and multimodal descriptions of spontaneous interaction as described in section 2.1. It allows the transcribers to distribute different descriptions onto different tiers according to which speaker and which level of description they belong to. In a musical score transcript, descriptions of sequential actions follow one another in a left-to-right reading direction, whereas simultaneous actions appear at the same horizontal position in a top-to-bottom reading direction. The number of tiers is, in principle, unlimited, and new tiers can be added (or reordered or deleted) at any point in the transcription process.

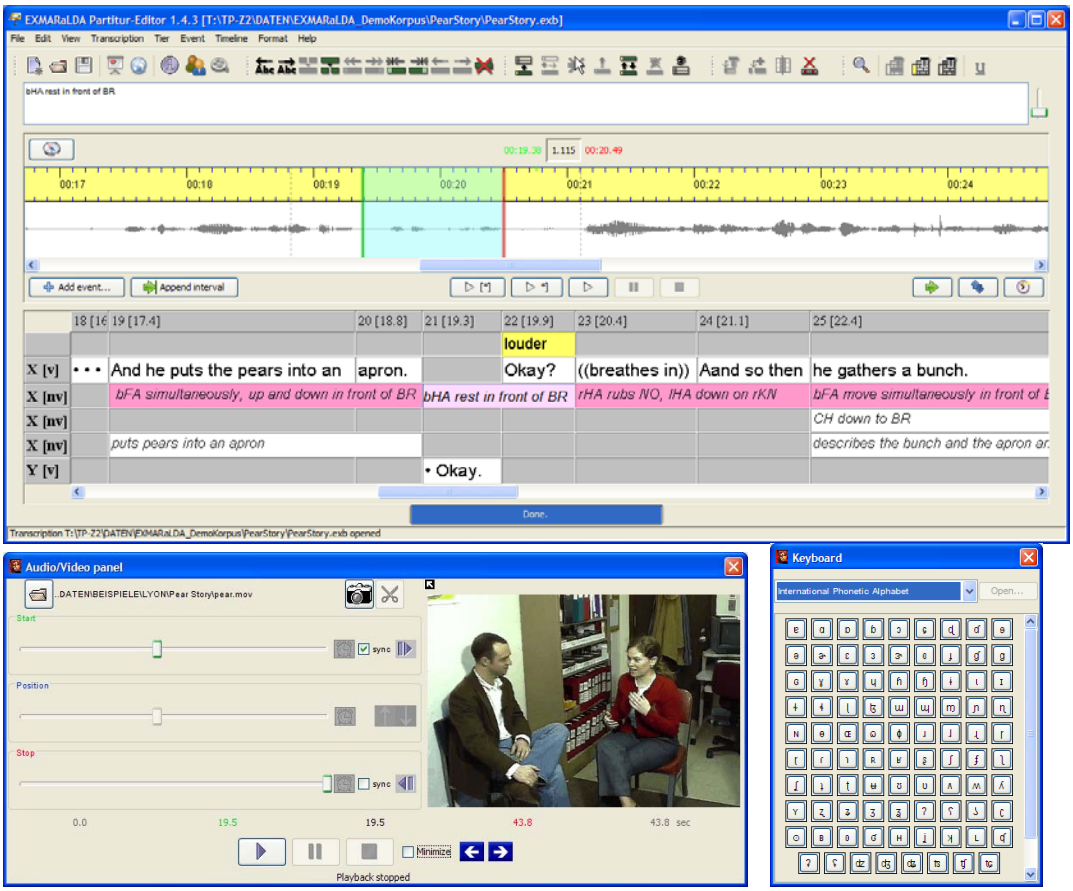

Figure 2: Screenshot of the EXMARaLDA Partitur-Editor with musical score interface (top), media player (bottom left) for synchronised media playback and virtual keyboard (bottom right) for input of special characters.

For output of transcriptions on screen or paper, the Partitur-Editor offers a number of functions for formatting tiers, wrapping musical scores to fit on a certain page width and integrating references to the underlying media signal. Different options for output formats, like HTML for presentation in a web browser, RTF for integration into MS Word documents or PDF for printing, are provided.

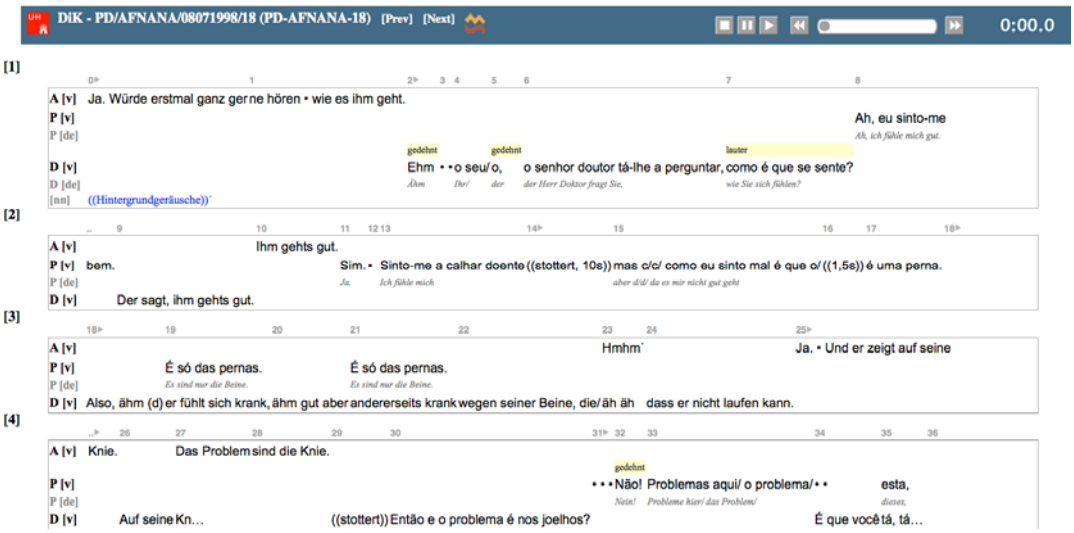

Figure 3: Visualisation of a transcription as a musical score, wrapped to fit on a certain page width

Regarding the output of transcriptions, users are, however, not restricted to the musical score layout. For certain types of analyses other presentation formats may be more helpful, and the Partitur-Editor enables the user to produce such presentation formats through the application of XSL stylesheets (a technology for transforming XML data). For example, figure 4 shows a list-like output format in which utterances are presented line-by-line in their temporal order (with links to the audio) alongside their translations.

1 [0]	**A**	Ja.	
2 [0.TIE1.1]		Würde erstmal ganz gerne hören • wie es ihm geht.	
3 [2]	**D**	Ehm • • o seu/ o, o senhor doutor tá-lhe a perguntar, como é que se sente?	Ähm Ihr/ der der Herr Doktor fragt Sie, wie Sie sich fühlen?
4 [8]	**P**	Ah, eu sinto-me bem.	Ah, ich fühle mich gut.
5 [9]	**D**	Der sagt, ihm gehts gut.	
6 [10]	**A**	Ihm gehts gut.	
7 [11]	**P**	Sim.	Ja.
8 [12]		• Sinto-me a calhar doente ((stottert, 10s)) mas c/c/ como eu sinto mal é que o/ ((1,5s)) é uma perna.	Ich fühle mich aber d/d/ da es mir nicht gut geht
9 [18]	**D**	Also, ähm (d)er fühlt sich krank, ähm gut aber andererseits krank wegen seiner Beine, die/ äh äh dass er nicht laufen kann.	

Figure 4: Visualisation of the same transcription as an utterance list with translations

3.3. *Creating and managing corpora and corpus meta-data: Corpus-Manager*

Whenever researchers work with more than one transcription, a need arises to organise recordings and transcriptions into corpora. Furthermore, as has been argued in section 2.2, it is often important for pragmatic research to be able to record detailed meta-data about interactions and participants. The relationship between interactions and partici-

pants and between recordings and transcriptions can be complex – interactions usually have more than one participant, one and the same participant can take part in different interactions, there may be several recordings of one and the same interaction and different transcriptions of one and the same recording. Because of this complexity, a tool is needed which helps researchers to bundle the different components of a corpus into a whole, to specify the relationships between them and to systematically describe them by appropriate meta-data sets.

In the EXMARaLDA system, this task is fulfilled by the Corpus-Manager (CoMa) software. CoMa allows the user to bundle transcriptions and recordings into corpora and to structure a corpus into communications (i.e. interactions) and speakers (i.e. participants). All components of a corpus can be described by a set of (arbitrary) meta-data attributes. Speakers and communications are independent units which can be assigned to one another. In that way, meta-data for speakers need not be duplicated when a speaker takes part in more than one communication.

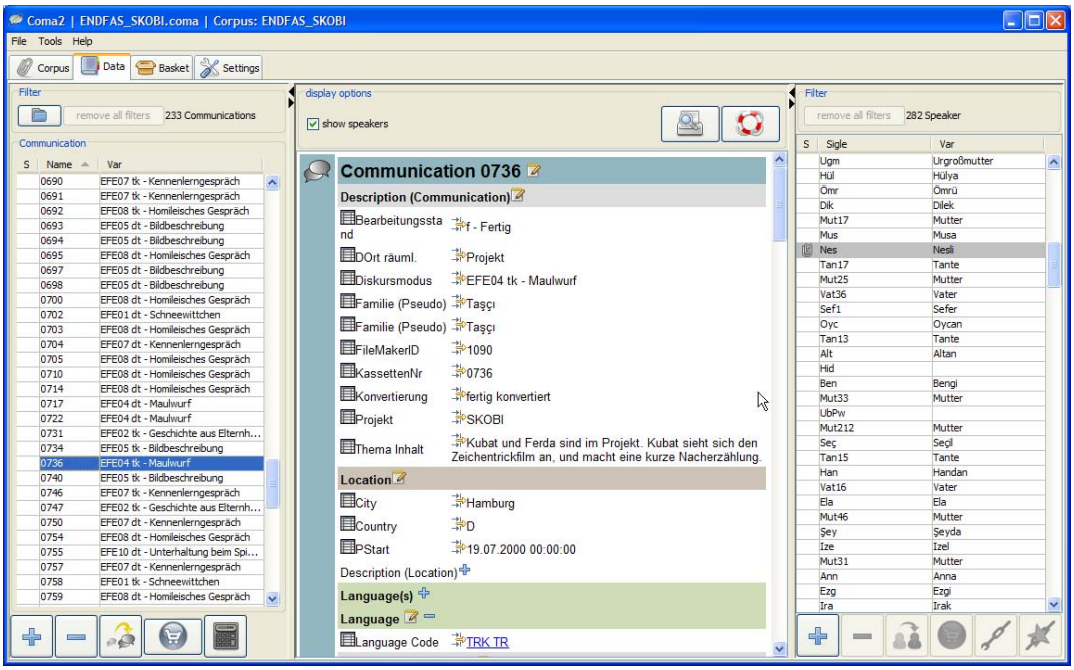

Figure 5: Screenshot of the Corpus Manager user interface with list of communications (left), communication meta-data (centre) and list of speakers (right)

Once a corpus has been created in that way, the Corpus Manager can also be used to carry out meta-data queries. Filters can be applied to create a subcorpus in which only those transcriptions are included whose corresponding meta-data sets have certain attributes. For instance, for a corpus of Turkish-German bilingual discourse, such a filter can select transcriptions of communications which took place in Turkey and which include a male bilingual speaker between the age of 8 and 12 years whose mother is also bilingual. The resulting subcorpus can then be saved separately and be used in EXAKT (see below).

3.4. *Querying and analysing corpora: EXAKT*

EXAKT (EXMARaLDA Analyse- und Konkordanztool) is a tool for querying transcription corpora for transcribed or annotated phenomena and for carrying out qualitative or quantitative analyses on the basis of such queries. The basic functionality of EXAKT is modelled after the classical corpus analysis instrument – a KWIC (keyword in context) concordancer. After having loaded a corpus compiled in the Corpus Manager, users can enter a search expression. Several types of search expressions are offered, the most common of which is a regular expression, i.e. a pattern specifying a string or a set of strings. For instance, the following are some regular expression typically used in a corpus query:

- [Tt]h(is|at|ose|ese) will match the words *this, that, those* and *these* and their capitalized variants.
- \bin[a-z]+abl[ey]\b will match words starting with *in* and ending in *able* or *ably* like *indisputable, indescribably, ineffable, indistinguishable* etc.
- (\b[A-Za-z]+\b){3,3}\? will match all sequences of three words followed by a question mark, i.e. the last three words of questions.

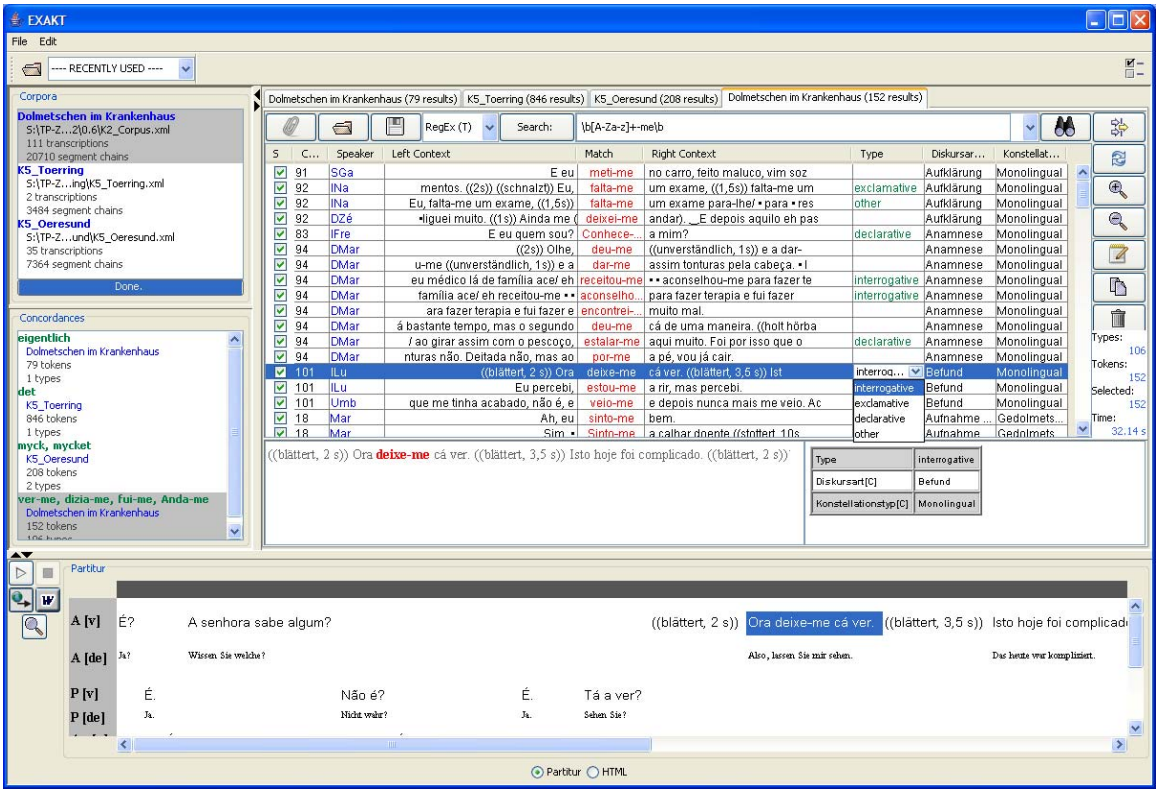

Figure 6: Screenshot of EXAKT with KWIC concordance table (centre) with additional columns for analysis and speaker/communication meta-data and musical score display (bottom).

As figure 6 demonstrates, the result of such a query is first presented as a keyword in context concordance, consisting of the matched expression itself with its immediately preceding and following context, typically the words uttered by the same speaker right

before and after the word(s) matched by the search expression. As in other concordancing tools, this result can then be sorted by the left or right context column in order to facilitate the discovery of context regularities.

As elaborated above, however, a pragmatically motivated corpus study can usually not restrict itself to this kind of context – more interactional context may be needed as well as situational context or ethnographic meta-data. For additional interactional context, EXAKT offers the possibility to display the corresponding part of a full musical score transcription (or the full transcription in some other layout) by double-clicking on any search result. Similarly, the corresponding part of the audio or video recording can be played back. In order to access meta-data about communications and speakers (as entered in CoMa), users can select arbitrary attributes to be displayed in additional columns of the KWIC table.

Frequently, the result of an automatic corpus query needs to be post-processed manually by the researcher. To support this task, EXAKT offers a number of filtering functionalities as well as the possibility to categorise search results with the help of one or more analysis columns.

When quantification is the aim of a corpus query, it is often not sufficient to simply count the total number of types or tokens in the corpus. Rather, quantification usually needs to take into account the afore-mentioned context information about speakers and communications in order to compare figures across different meta-data attributes. This typically requires summarizing search results with certain attributes into groups and then quantifying according to this grouping. To support this kind of analysis, EXAKT offers the possibility to apply XSL stylesheets to a given search result. Figure 7 shows a search result for German wh-Words in which this technique has been used to group results first by speaker, then by speaker age, and finally by word type.

Bur2 (15)

Age	Count	Types
9;1	9	Wer (3) wenig (1) wer (4) werd (1)
9;7	1	wenn (1)
11;1	5	wenn (5)

Dil (11)

Age	Count	Types
9;4	11	Wel (1) Wer (3) welche (1) welchen (1) welcher (1) welches (1) wem (1) wenn (2)

Fik (19)

Age	Count	Types
6;4	3	weem (1) wenn (1) wer (1)
6;5	10	Wer (2) welche (4) wenn (2) wer (2)
6;9	1	Wer (1)
6;10	4	Wer (2) wenn (1) wer (1)

Figure 7: Quantification of a grouped search result.

4. EXMARaLDA corpora

At the Research Centre on Multilingualism at the University of Hamburg where EXMARaLDA is developed, a number of research projects use EXMARaLDA in their work with spoken language corpora. Basically, these projects come from two different theoretical backgrounds. On the one hand, there are researchers studying bilingual first language acquisition in a generativist framework. Since they are primarily interested in the syntactic or phonological properties of individual children's utterances, their transcriptions and corpora are usually less comprehensive in terms of the number and detail of description levels. On the other hand, there are several projects studying multilingual communication in a discourse analytic framework based on the school of functional pragmatics. It is these projects that the observations about corpus-based pragmatics in section 2 come from, and we will consequently concentrate on three of their corpora in the remainder of this section.

4.1. *The corpus "Scandinavian Semicommunication"*

The aim of the project "Semi-Communication and Receptive Multilingualism in Scandinavia" (headed by Kurt Braunmüller, see also Braunmüller 2000) was to investigate inter-Nordic communication or, more specifically, the mutual understanding between speakers of Danish, Swedish, and Norwegian. To this end, a corpus of spoken language was compiled using recordings from different domains in which such semi-communication customarily occurs. More specifically, the corpus consists of the following six subcorpora:

- NUAS: Recordings and transcriptions from three conferences of the Nordic Association of University Administrators (NUAS), an organisation fostering the establishment of networks between the Nordic universities at different administrative levels.
- Öresund Direkt: Recordings and transcriptions of the radio program "Öresund Direkt", broadcast jointly by a Swedish and a Danish broadcasting station in the Öresund region.
- Radio recordings: Recordings and transcriptions of other radio programs in which semi-communication occurred.
- German school: Recordings and transcriptions from school lessons at a German school for Swedish, Danish and Norwegian children.
- Danish school: Recordings and transcriptions from school lessons at a Danish school while a group of Norwegian children was visiting.
- University courses: Recordings and transcriptions from language courses at a Swedish university where Danish native speakers taught Danish to Swedish native speakers.

Altogether the corpus consists of around 90 hours of audio material, around 50% of which has been transcribed. There are a total of 74 transcriptions, amounting to 269,945 transcribed words.

	Transcriptions	Total words	Recordings	Total duration
NUAS	18	34,728	105	38:56:01
Öresund Direkt	35	147,325	41	28:58:41
Radio Recordings	3	1,939	3	0:12:28
German school	3	6,252	15	6:00:10
Danish school	2	19,759	2	1:28:36
University courses	13	59,942	17	13:02:29
Total	**74**	**269,945**	**183**	**88:38:25**

Transcriptions were done according to the HIAT transcription system (Rehbein et al. 2004), transcribing verbal behaviour in a modified orthography ('literary transcription' as termed by Ehlich 1993) and taking note of pauses and other non-phonological phenomena. In some places, code switches were annotated and transcribers included comments on unusual linguistic phenomena. Figure 8 is an excerpt of a typical transcription from this corpus.

[28]
[nsp] ((0,6s))
Dm1 [v] ((holt Luft,0,5s)) Er det noget der kommer til udtryk sådan i dagligdagen, eller
Sw9 [v] så.

[29]
Dm1 [v] bare noget man kan se på medarbejderne? ((0,3s)) At danskerne er kraftigere, eller…
Sw9 [v] Nej!

[30]
[nsp] ((0,5s)) ((0,8s))
Dm1 [v] Nej.
Sw9 [v] Nej, nej. Det är de inte, inte så, men jag tror nog att vi påpekar för

[31]
Sw9 [v] varandra, "ät lite mer sallad" och ja…

Figure 8: An excerpt from a transcription of the "Öresund Direkt" corpus. The Danish speaker (Dm1) and the Swedish speaker (Sw9) each user their mother tongue in communicating with each other.

4.2. *The corpus "Interpreting in hospitals"*

The project "Interpreting in hospitals" (headed by Kristin Bührig and Bernd Meyer, see also Meyer 2000) investigated doctor-patient communication mediated by non-trained interpreters, for example relatives of the patient or bilingual staff members. To this end, a spoken language corpus was compiled using recordings from doctor-patient interactions in different monolingual and bilingual language settings. More specifically, the corpus consists of the following five subcorpora:

- German monolingual: Recordings and transcriptions of communication between a German speaking doctor and a German speaking patient
- Portuguese monolingual: Recordings and transcriptions of communication between a Portuguese speaking doctor and a Portuguese speaking patient

- Portuguese bilingual: Recordings and transcriptions of communication between a German speaking doctor and a Portuguese speaking patient, mediated by a Portuguese/German bilingual interpreter
- Turkish monolingual: Recordings and transcriptions of communication between a Turkish speaking doctor and a Turkish speaking patient
- Turkish bilingual: Recordings and transcriptions of communication between a German speaking doctor and a Turkish speaking patient, mediated by a Turkish /German bilingual interpreter

	Transcriptions	**Total words**	**Recordings**	**Total duration**
German mono	14	17,467	14	2:49:50
Portuguese mono	24	31,404	21	3:48:45
Portuguese bi	38	54,495	25	6:47:06
Turkish mono	13	16,070	12	2:39:59
Turkish bi	23	51,404	20	6:55:38
Total	**112**	**170,840**	**92**	**25:01:18**

Altogether the corpus consists of around 25 hours of audio material, almost all of which has been transcribed. There are a total of 112 transcriptions, amounting to 170,840 transcribed words.

Transcriptions were done according to the HIAT transcription system transcribing verbal behaviour in a modified orthography and taking note of pauses and other non-phonological phenomena. Suprasegmental characteristics (prosody) and accentuation were annotated in separate tiers. For all Portuguese and Turkish utterances, an interlinear translation was provided to facilitate analysis and presentation of the data. Figure 9 is an excerpt of a typical transcription from this corpus.

[4]

A [v]	jetzt nich… So! Ich möchte/ Sie sind ja schon an der Hüfte operiert
P [v]	((unverständlich)).

[5]

A [v]	worden, ((1s)) vor drei Wochen, ((2s))
D [v]	Já foi/ eh você já foi operado
D [de]	*Sie wurden/ äh Sie wurden schon einmal*
P [v]	Ai, eu, agora eu, eu não sei o que responder.
P [de]	*Nun, ich, jetzt, ich, ich weiß nicht, was ich antworten soll.*

[6]

A [v]	vor drei Wochen.
D [v]	uma vez à esquerda. Eh noo/ que eh/ de/ está-lhe a
D [de]	*links operiert.* *Äh am/ dass äh/ von/ sie sagt Ihnen,*
P [v]	No dia/ eh no dia cinco. No
P [de]	*Am fünf/ äh am fünften. Am*

Figure 9: An excerpt from a transcription of a Portuguese/German bilingual doctor-patient conversation. The conversation between the German speaking doctor (A) and the Portuguese speaking patient (P) is

mediated by a Portuguese/German bilingual interpreter (D). Portuguese utterances have been provided with an interlinear translation into German (D [de] and P [de]).

4.3. *The corpus "Turkish-German bilingualism"*

The projects ENDFAS and SKOBI (headed by Jochen Rehbein, see also Baumgarten et al. 2007) investigated various aspects of bilingual language acquisition of Turkish/German bilingual children, comparing them to monolingual acquisition settings. To this end, a corpus was compiled using recordings elicited by so-called evocative field experiments, a method developed in the projects to create quasi-natural constellations coming as close as possible to everyday communicative practice and at the same time enabling repeated, comparable recordings. Examples of such field experiments are story retellings, verbalisations of cartoons or homileic talks about friends, television, toys or illnesses.

The editing of the corpus is about 80% complete. In its current state, the corpus comprises around 700 transcriptions, amounting to around 700,000 transcribed words. It is estimated that these figures will increase to around 1000 transcriptions with approximately 1 million words when editing is complete.

Transcriptions were done according to the HIAT transcription system transcribing verbal behaviour in a modified orthography and taking note of pauses and other non-phonological phenomena. Suprasegmental characteristics (prosody) and accentuation were annotated in separate tiers. For all Turkish utterances, an interlinear translation was provided to facilitate analysis and presentation of the data. Figure 10 is an excerpt of a typical transcription from this corpus.

[1]

Kub		Nein.	
Fer	Lütfen başından sona kadar anlatır mısın bana?	Do…	Güzel oturur
Fer [de]	Würdest du es mir bitte von Anfang bis zum Ende erzählen?		Würdest du dich

[2]

Kub	• • • Nasıl?			
Kub [de]	Wie?			Wie?
Fer	musun oraya?	Bak dizimi çok acıttın Kubat!		
Fer [de]	dort richtig hinsetzen?	Schau, du hast meinem Knie sehr weh getan, Kubat!		

[3]

Kub	Nasıl?	Hab ich. ((1,5s))	Daa ist doch	kein Fernsehen.
Kub [de]				
Fer	((unv.)).		((unverstdl.)).	((güler))
Fer [de]				((lacht))

[4]

Fer	((1s)) Hadi anlatır mısın bana şimdi?		
Fer [de]	Los, würdest du mir jetzt bitte erzählen?		
Nes		((1s)) Yer misiniz?	
Nes [de]		Möchtet ihr essen?	

Figure 10: An excerpt from a transcription of a story retelling (evocative field experiment 04). The bilingual child (Kub) is asked to retell a TV cartoon he has just seen. The retelling is meant to be done in Turkish, but since the other participants (Fer and Nes) are also bilingual, code switching occurs. Turkish utterances have been provided with an interlinear translation into German (Kub [de], Fer [de] and Nes [de]).

4.4. *Corpus edition and publication*

The three corpora mentioned above, like the corpora listed in the following section, were (or are) compiled using the EXMARaLDA Partitur-Editor and the Corpus-Manager. Partly, the transcriptions had been originally created with other tools (HIAT-DOS, syncWriter or Praat) and were only later converted to the EXMARaLDA format.[4]

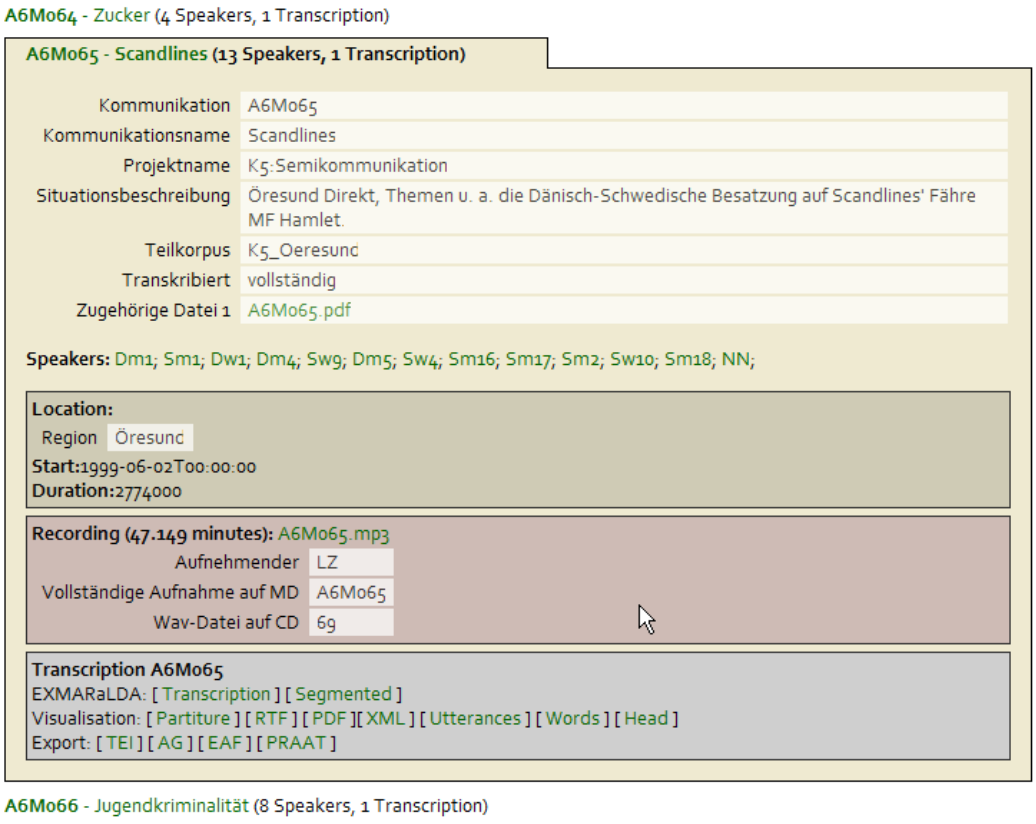

Figure 11: An HTML visualisation of a piece of communication meta-data with links to different archiving and presentation files of the corresponding transcriptions.

After compilation is complete, each corpus is subjected to a final editing process. This includes a final check for errors or inconsistencies in the data, supplementing missing

[4] Corpus compilation had begun and, in some cases was almost completed, before the development of EXMARaLDA started. In part, EXMARaLDA's objective was also to "rescue" transcription data from tools whose formats are problematic in terms of data reuse and archiving (see also Schmidt/Bennöhr 2008). Newer projects now usually use the Partitur-Editor or compatible tools from the start to create their transcriptions.

meta-data and synchronisation, and the packaging of the corpus into an archivable and publishable form. For each communication in the corpus, three folders are created: One folder contains the actual EXMARaLDA data and the digitized media file (in MP3 format for easy distribution, in WAV format for archiving). This is in principle sufficient to work with the corpus, using the EXMARaLDA tools described above. Another folder contains different presentation files such as a musical score visualisation in HTML, PDF and RTF, and other visualisations of the transcription and the meta-data in HTML. The HTML files are hyperlinked between one another and contain links to the underlying media file where appropriate. With this hyperlinked HTML structure, corpus users are offered a convenient way of exploring the corpus with the help of a standard internet browser.

The third folder, finally, contains the transcription data in other widely-used archivable formats. For instance, for each transcription a version is provided which represents the data not in the EXMARaLDA XML format, but in an XML format adhering to the TEI guidelines for transcriptions of speech (sees Schmidt 2005). This is intended to further facilitate data exchange and reuse.

Corpora which have been completely edited in this way (presently the corpora presented in sections 4.1 and 4.2, for all others[5], the editing process is ongoing) will be made publicly available as far as copyright and privacy protection regulations allow this.

5. Conclusion and outlook

In this paper, we have introduced a system for the computer-assisted creation and analysis of spoken language data and presented three corpora which were compiled with the help of this system. In concluding, we would like to summarise once again the main reasons why we think that systems like EXMARaLDA can make an important contribution to the field of corpus-based pragmatics:

1) Corpus-based pragmatics deals with data structures which are more complex than those of 'classical' corpus linguistic analyses. EXMARaLDA supports this kind of complexity through an adequate data model allowing the integration of different levels of linguistic description and different types of context data.
2) Pragmatics has an explorative approach to corpora, corpus-based pragmatics is often actually corpus-driven pragmatics. EXMARaLDA supports this corpus-driven approach through software tools offering flexible data visualisation and permitting a two-way interaction between the data and its analysis.

[5] Besides the corpora presented in sections 4.1 to 4.3, several other spoken language corpora are presently being compiled and edited at the Research Centre on Multilingualism. Most importantly, these comprise a number of bilingual language acquisition corpora consisting of longitudinal interview data with bilingual (French/German, Spanish/German, Portuguese/German, Spanish/Basque) children. In addition, there is a corpus with Portuguese/German simultaneous interpreting data, a corpus of data from Faroese/Danish bilinguals, a corpus for the study of specific language impairment (SLI) in Turkish/German bilinguals, and a corpus of present-day Catalan. Together, these corpora cover more than 1,500 hours of audio or video recordings corresponding to an estimated 2.5 million transcribed words. Completed corpora are available from http://www.exmaralda.org/corpora/en_sfbkorpora.html.

3) Corpus-based pragmatics profits from the researchers' ability to reuse, exchange and archive valuable language corpora. EXMARaLDA supports the sustainability of linguistic data through the use of open standards.

Six years after the beginning of its development, EXMARaLDA is now a stable system used by a great number of researchers from different backgrounds. Development and maintenance of the EXMARaLDA software tools and of spoken language corpora are ongoing, as is the constant exchange with other projects and initiatives with a similar objective. In the long run, we hope that EXMARaLDA will thus make a substantial contribution to the goal of enabling linguists to work efficiently and productively with large bodies of empirical data.

References

Baumgarten, Nicole, Annette Herkenrath, Thomas Schmidt, Kai Wörner, and Ludger Zeevaert (2007) Studying connectivity with the help of computer-readable corpora: Some exemplary analyses from modern and historical, written and spoken corpora. In Jochen Rehbein, Christiane Hohenstein, and Lukas Pietsch (eds.), *Connectivity in Grammar and Discourse*. Amsterdam: Benjamins Publishing Company, pp. 259-289.

Braunmüller, Kurt (2000) Semikommunikation in phatischen Dialogen. In Bernd Meyer, and Notis Toufexis (eds.), Text/Diskurs, Oralität/Literalität unter dem Aspekt der mehrsprachigen Kommunikation. Beiträge zum Workshop, Methodologie und Datenanalyse'. *Working Papers in Multilingualism*, Series B (11). Hamburg, pp. 101-114.

Bird, Steven, and Mark Liberman (2001) A formal framework for linguistic annotation. *Speech Communication* 33.1,2: 23-60.

Bird, Steven, and Gary Simons (2003) Seven dimensions of portability for language documentation and description. *Language* 79: 557-582.

Deppermann, Arnulf (2000) Ethnographische Gesprächsanalyse: Zu Nutzen und Notwendigkeit von Ethnographie für die Konversationsanalyse. *Gesprächsforschung* 1: 96-124.

Edwards, Jane (1993) Principles and contrasting systems of discourse transcription. In Jane Edwards, and Martin Lampert (eds.), *Talking Data – Transcription and Coding in Discourse Research*. Hillsdale: Erlbaum, pp. 3-31.

Ehlich, Konrad, and Jochen Rehbein (1976) Halbinterpretative Arbeitstranskriptionen (HIAT). *Linguistische Berichte* 45: 21-41.

Ehlich, Konrad (1993) HIAT - a transcription system for discourse data. In Jane Edwards, and Martin Lampert (eds.), *Talking Data – Transcription and Coding in Discourse Research*. Hillsdale: Erlbaum, pp. 123-148.

Isard, Amy, David McKelvie, and Henry Thompson (1998) Towards a minimal standard for dialogue transcripts: A New SGML Architecture for the HCRC Map Task Corpus. *Proceedings of the 5th International Conference on Spoken Language Processing*. Sydney.

Meyer, Bernd (2000) Zur Analyse gedolmetschter Arzt-Patienten-Kommunikation im Krankenhaus. In Bernd Meyer, and Notis Toufexis (eds.), Text/Diskurs, Oralität/Literalität unter dem Aspekt der mehrsprachigen Kommunikation. Beiträge zum Workshop ‚Methodologie und Datenanalyse'. *Working Papers in Multilingualism*, Series B (11). Hamburg, pp. 45-53.

Ochs, Elinor (1979) Transcription as theory. In Elinor Ochs, and Bambi Schieffelin (eds.), *Developmental Pragmatics*. New York, San Francisco, London: Academic Press, pp. 43-72.

Rehbein, Jochen, Wilhelm Grießhaber, Petra Löning, Marion Hartung, and Kristin Bührig (1993) Manual für das computergestützte Transkribieren mit dem Programm syncWRITER nach dem Verfahren der Halbinterpretativen Arbeitstranskriptionen *(HIAT).* Hamburg: Germanisches Seminar, Universität Hamburg.

Rehbein, Jochen, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath (2004) Handbuch für das computergestützte Transkribieren nach HIAT. *Working Papers in Multilingualism*, Series B (56). Hamburg.

Schmidt, Thomas (2005) Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. *Working Papers in Multilingualism*, Series B (62). Hamburg.

Schmidt, Thomas, and Jasmine Bennöhr (2008) Rescuing legacy data. *Language Documentation & Conservation* 2.1: 109-129.

Teubert, Wolfgang (2005) My version of corpus linguistics. *International Journal of Corpus Linguistics* 1/2005.